Unit 8.5 - Application of Statistical Inference to Capstone Project

Summary:

Continuing answering the questions proposed for investigation, I applied techniques in statistical inference for categorical variables. I applied the Chi-Square test and used the Bootstrap method to infer visually if a subset of a group may have a differentiable mean from the overall average.

Chi-Square Test:

Since this project is a classification problem and the target variable and the most of the independent variables are categorical, I felt that the chi-square test for between each independent variable and the dependent variable.

The null hypothesis (H0) is that there is no difference between the proportion of the zip code that has caravan insurance for each variable than the overall population, without discerning specific variable(s).

The alternative hypothesis (H1) is that there is a difference.

The formal expression is:

H0: U\_xi = U\_y

H1: U\_xi != U\_y

I leverage the Sci-Kit Learn built in Chi-Square Contingency table to get the Chi Square statistic and calculate the P-value from a chi-square distribution. I do it one variable against the target variable at a time using a loop.

Below is the chi-square test result for the first 9 variables, out of total 85.

|  |  |  |
| --- | --- | --- |
| **variable name** | **chi square stat** | **p-value** |
| Customer Subtype | 124.81 | 0.001 |
| Number of houses | 3.46 | 0.9999 |
| Avg size household | 9.33 | 0.5015 |
| Avg age | 3.29 | 0.9931 |
| Customer main type | 88.66 | 0 |
| Roman catholic | 9.12 | 0.9815 |
| Protestant | 22.4 | 0.3192 |
| Other religion | 11.74 | 0.4671 |
| No religion | 19.4 | 0.4959 |

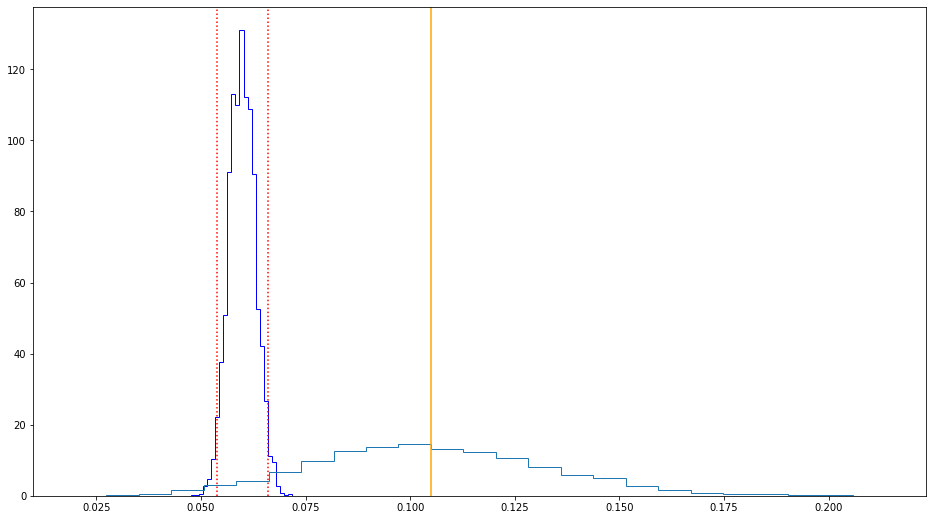
From that table, we may infer that the variable ‘Customer Subtype’, for example, can be statistically significant, because the p-value is significantly smaller than 5%.

Bootstrap Inference

Beside from the Chi-Square test, I also applied bootstrap inference as my second approach. The main idea here is to compare the overall mean or proportion of the sample population that has caravan insurance.

As an example, here is a visual example.

1



The dark blue histogram is the bootstrap distribution of the mean and the red vertical lines are the boundaries for the 95% confidence intervals of the mean. The orange line is the mean of the subset of the first independent variable with classification as 1. The first variable has 42 different classes. The light blue line is the bootstrap distribution of the subset. From this chart, we may conclude that the mean of this subset seems significant, though the variance may be quite high.

Conclusion

I hope presenting the two approaches allowed me to see better what part of the data seems potentially predictive as I move forward in constructing and testing different models.